ERRONEOUS AND EXCESSIVE TREATMENT ANOMALY DETECTION BASED ON UNSUPERVISED LEARNING

Shisheng ZHU[1,*], Chaoqing LI[2], Muwan LIU[3], Wei LIU[4]

To provide an effective way to optimize the diagnosis and treatment process, in this research, our model of erroneous medical diagnosis is developed from the association rule learning by Apriori algorithm in order to corelate diseases and symptoms, and further expanded with an innovative numerical quantification method of prescription therapeutic function (called "Treatment Weight Effect Approach"). This is the first time an optimal expert prescription can be discovered from a large scale set of similar prescriptions without requiring explicit labels of training data. In this study, the idea of Hamming distance in K-Modes clustering algorithm is used to calculate the similarity of drugs, and then new prescriptions are compared with the best expert prescription to detect excess treatment. Additional experiments are conducted to verify the feasibility of the program with diagnosis and prescription treatment data. The testing results have shown that the model is successful in providing detection of erroneous and excessive treatments from the existing prescriptions and treatments. We address the critical issue of Treatment Anomalies due to Erroneous and Excessive Medications. We have overcome the traditional medical anomaly detection methods using data mining, Bayesian classification, decision trees and other supervised learning detection methods that were limited in scale and lacked data labels.

Keywords: Erroneous and excessive treatment, association rule, "Treatment Weight Effect Approach", Hamming distance, similarity

1. Introduction

Medical and health care technology has continued to advance and gradually become matured in many areas. However, the anomalies of erroneous and excessive treatment still cause- huge waste of social medical resources and -reduce- the patients' satisfaction with medical treatment. Erroneous treatment refers to a medical activity in which a doctor gives the wrong treatment to a

[1] Department of Computer Science, Shantou University, Shantou 515063, China, E-mail: sszhu@stu.edu.cn

[2] Department of Computer Science, Shantou University, Shantou 515063, China, E-mail: 17cqli1@stu.edu.cn

[3] Department of Computer Science, Shantou University, Shantou 515063, China, E-mail: 1530463576@qq.com

[4] School of Science & Technology, G.G.C, Lawrenceville 30043, USA, E-mail: wei-liu@att.net, *Corresponding author: Shisheng Zhu, E-mail: sszhu@stu.edu.cn

patient, including prescribing the wrong medicine; excessive treatment refers to the irregular and incorrect medical treatment behaviors carried out by doctors out of patients' actual conditions in the medical process, including over-medication, over-diagnosis and so on [1][2]. Erroneous and excessive treatment is ubiquitous worldwide. The International Organization for Migration report estimated that 44,000-98,000 people died each year from medical malpractice in the United States. According to the survey, the incidence of adverse events among hospital patients around the world was relatively high due to erroneous treatment [3]. According to reports provided by many online hospitals in China by 2015, the number of erroneous treatment cases caused by medication errors has reached 11,792 [4]. In the case of the excessive treatments, such as the excessive X-ray examination of the female patients in the United States [5], excessive examination and overdose have exacerbated the patients' cost burden and psychological shadow. In addition, inconsistencies between pathology and clinical treatment decisions have led to excessive treatment of benign diseases [6] in several children in the United States.

While erroneous and excessive treatments are mainly caused by the immature medical technology as well as the imperfect laws and regulations, the under-quality of some medical workers can also cause those due to human errors. Furthermore, reducing medical errors and optimizing the medical diagnosis and treatment model is also particularly urgent in order to ease the tense relationship between doctors and patients.

In order to improve the quality of medical treatment for patients, many data mining methods have been applied to the abnormal detection and resolution of medical diagnosis and treatment. One of the solutions, supervised learning methods are of task-driven detection, applies decision trees and support vector machine (SVM) classification algorithms to detect erroneous and excessive treatment anomalies. In this type of application, the classification method requires labeled medical data for training [7], but most medical data are unlabeled. To overcome the issue of lacking data labels, one can apply unsupervised data mining methods. The unsupervised approach, which is detected in a data-driven manner, can effectively detect abnormal medical information during the medical diagnosis and treatment process [8], thus finding erroneous and excessive treatment. Therefore, this study uses the Apriori unsupervised learning algorithm and the Hamming distance method. By using the idea of Hamming distance in K-Modes clustering algorithm, this study deals with classified attribute medical data, uses the heterogeneity to measure the similarity between drugs using different methods, and identifies anomalies of excessive treatment in a short time. At the same time, the Apriori algorithm is used to mine the association rules of diseases and symptoms, so as to identify erroneous treatment.

In this paper, the study on erroneous and excessive treatment anomalies based on unsupervised learning is carried out, and the research program of erroneous and excessive treatments based on association rule and Hamming distance is presented. This program could transform the traditional medical model and establish a patient-centered medical model. In the process of medical abnormality discovery, this study also uses the optimized disease diagnostic model, in which medical abnormal reasoning mechanism and rules can be continuously accumulated and constructed. Combing the unsupervised self-learning with optimized treatment model enables even stronger self-accumulation and self-learning ability, and further strengthens the judgment accuracy of erroneous treatment. For the judgment of excessive treatment, the similar-drug-comparison model is used to compare the similarity of the drugs themselves in the new prescription and the best prescription, and then the similarity between the prescriptions at the treatment level is compared, which provides certain authenticity. As a result, the resulting program allows regulation of the medical treatment process and provides patients with a safe medical care model.

The main contributions of this study are as follows:

(1) A new disease diagnosis and judgment model based on the association rule mining method of Apriori algorithm was built. The model can calculate the correlation between disease and symptoms as well as between diseases, so as to establish a set of rule knowledge for diagnosis of anomalies.

(2) “Treatment Weight Effect Approach” was developed in an innovative manner, which quantified the treatment effect of the drug therapeutic function of the prescription on the symptoms, in order to find the best expert prescription from the similar prescription set.

(3) The Hamming distance method was used to analyze the similar prescription issues, the similarities between drugs corresponding to new prescription and expert prescription are judged by calculating the distance of drug therapeutic functions, and then the two prescriptions were compared for obtaining the similarities.

(4) Ultimately, regulatory recommendations could be formed from the existence of prescription treatment and the cost of excessive treatment.

The structure of this paper is as follows. The first section introduces the research background. The second section deals with the summary of related work, the research framework and research methods of erroneous and excessive treatment. Section 3 describes the experimental process and results discussion.

2. Materials and methods

This part includes related work, research framework and research methods.

2.1 Related work

Abnormal medical problems such as erroneous and excessive treatments have attracted the attention of scholars at home and abroad. Gelatti et al. [9]

discussed a high-performance anomaly detection method that combined medical anomaly detection mining techniques with temporal abstraction techniques, which supported the diagnosis and monitoring of intensive care units. Since this method did not discuss the detection of numerical data and classification attributes data, it was unfavorable for the monitoring of unstructured and semi-structured medical data errors. Tago et al. [10] proposed an improved abnormal treatment method that considered the patients' potential factors and status, and statistically detected medical data by taking advantage of the Hotlin theory, yet it had to examine the effectiveness with labeled medical data. Zhang et al. [11] propose a new unsupervised method to detect abnormal cardiac signals by similarity comparison. The program is mainly for ECG (Electrocardiogram) detection, helping doctors to conduct medical diagnosis correctly and avoid wrong treatment. Nevertheless, the applicability of the program remains to be studied. In order to solve the multiple attributes of medical health data and the coexistence of continuous discrete types, Liu et al. [12] studied the classification rule mining method on mixed data and improved the classification efficiency of rough set knowledge discovery method on medical health data. Song et al. [13] put forward a semi-supervised hybrid anomaly detection model for high-dimensional data, which improved detection accuracy and reduced computational complexity. For supervised learning classification methods or semi-supervised models, the unsupervised learning approach is more advantageous for the characteristics of medical data itself.

Rush et al. [14] discuss the application and case of machine learning applied to medical data, and propose to strictly evaluate the application of machine learning in the medical field, thus avoiding erroneous and excessive treatment to a certain extent, and improving disease detection, clinical decision-making, etc. The unsupervised learning method belongs to the field of machine learning. It can start from the nature of unmarked data without prior knowledge, summarize data characteristics and solve various problems. Goldstein et al. [15] discussed 19 different unsupervised anomaly detection algorithms from the perspective of anomaly detection performance, calculation amount, parameter setting and so on. The study outlined the characteristics of anomaly detection and suggested recommendations for anomaly detection in real-life situations. Amor et al. [16] propose a new anomaly detection program to detect erroneous measurements and to better distinguish between faults and real medical events. Feedback on the cause of medical testing needs further research to optimize diagnostic rules. The association rule is a type of unsupervised learning method, so the corresponding association rule can be designed to study the correlation between anomaly detection and possible causes. Given the limitations of previous methods, the association rules of the Apriori algorithm is used to discover the erroneous medical problems, to establish a disease diagnosis model, and to identity the abnormal treatment issues. At the same time, "Treatment Weight

Approach" and Hamming distance are adopted to address similarity issues and avoid excessive treatment. The results show that the application of the two methods can effectively improve the traditional medical treatment methods.

2.2 Research framework of erroneous and excessive treatment problems

Erroneous and excessive treatment anomalies are related to medical diagnosis and treatment activities, thus we need to research on those related activities as well. In order to cover the patient treatment process, we design the research framework of erroneous and excessive treatment anomalies, as shown in Fig. 1.

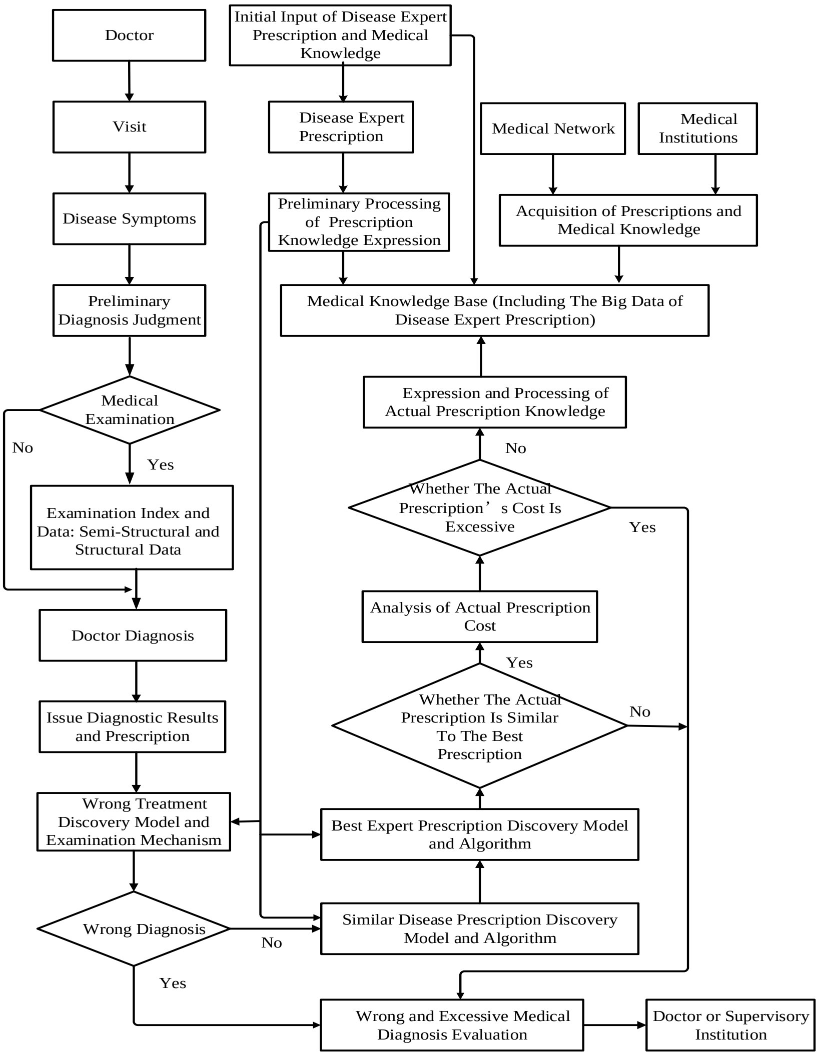

Doctor
Initial Input of Disease Expert Prescription and Medical Knowledge
Visit
Disease Expert Prescription
Medical Network
Medical Institutions
Disease Symptoms
Preliminary Processing of Prescription Knowledge Expression
Acquisition of Prescriptions and Medical Knowledge
Preliminary Diagnosis Judgment
Medical Knowledge Base (Including The Big Data of Disease Expert Prescription)
Medical Examination
Expression and Processing of Actual Prescription Knowledge
No
Yes
No
Examination Index and Data: Semi-Structural and Structural Data
Whether The Actual Prescription's Cost Is Excessive
Yes
Doctor Diagnosis
Analysis of Actual Prescription Cost
Yes
Issue Diagnostic Results and Prescription
Whether The Actual Prescription Is Similar To The Best Prescription
No
Wrong Treatment Discovery Model and Examination Mechanism
Best Expert Prescription Discovery Model and Algorithm
Wrong Diagnosis
No
Similar Disease Prescription Discovery Model and Algorithm
Yes
Wrong and Excessive Medical Diagnosis Evaluation
Doctor or Supervisory Institution

Fig. 1 Research framework of erroneous and excessive treatment problems

The framework starts from patient visit and doctor consultation. Doctors make initial diagnosis and judgment based on disease symptoms, and issues diagnosis results and drug prescriptions among them. The data types from diagnostic result are more than just numerical data, they could include binary or discrete types. The data type are running through the association mechanism and the erroneous treatment discovery model, during which we use association mining to analyze the diagnosis results and drug prescriptions, so as to judge whether there is a wrong treatment in the disease diagnosis. In this study, Hamming distance knowledge is used to analyze the similarity between the correct treatment prescription and best prescription, in order to exclude excessive treatment anomalies in prescriptions. Next we also judge whether new prescriptions are more expensive or not than the best prescription. The process ultimately evaluates the results of erroneous and excessive treatment anomalies for the regulatory purposes.

Throughout the process, the medical treatment knowledge base supports the discovery and examination of erroneous and excessive treatments. The knowledge base is formed by accumulation of drug knowledge and expert knowledge together with the drug prescription after each diagnosis. The knowledge base is divided into three categories: drug knowledge base, diagnosis and treatment knowledge base, and expert prescription knowledge base. The drug knowledge base contains the many-to-many relationship between drugs and therapeutic functions and the many-to-one relationship between drug indication functions and adaptive symptoms. The diagnosis and treatment knowledge base contains information about anti-retroviral drugs for various diseases, a mixture of different drugs that cannot be combined. The diagnosis and treatment knowledge base also contains the disease rule base on how the symptoms described by the patient can be used to infer the disease condition. The disease rule base is in turn trained by the disease diagnostic model. The expert prescription knowledge base contains the examination items and costs corresponding to the disease indicated by the expert prescription, as well as the better drugs for the disease and the corresponding relationship between the disease and the symptoms. The functions of knowledge base accumulation are implemented through the collaborative diagnosis and treatment platform [17] established by our research team. The platform considers the main challenges in the e-health interconnection network services [18], so as to better accumulate knowledge in the Internet services. While it is based on the collaborative diagnosis and treatment platform, this paper will focus on the research program for erroneous and excessive treatment anomalies. The new tool sets are unsupervised leaning via Hamming distance and Apriori algorithm. Additional extensions are also developed as described in detail in the next section.

2.3 Research program for medical anomalies of unsupervised learning

In this part, erroneous and excessive treatment discovery methods are analyzed.

2.3.1 Erroneous Treatment Discovery Methods

There may be erroneous medical problem in the prescriptions given by doctors, so it is necessary to explore the method of finding erroneous medical diagnosis. Therefore, this part first formalizes the disease and disease symptoms, and then constructs the erroneous medical discovery model.

(1) Formal abstract representation

The expert disease prescription knowledge base contains a large amount of information on diagnosis and treatment, and the data provided by such information can be used as a training set for erroneous medical discovery model. Due to the huge amount and complexity of medical data, the data mining work cannot be done well. Therefore, the medical treatment, disease and symptoms were formally abstracted, and the medical data table was simulated, as shown in Table 1. Among them, the visit number is equivalent to the transaction item in the Apriori algorithm, representing a visit to the patient; the symptom code indicates the symptom described by the patient during the visit; the disease code indicates that the doctor diagnosed the patient’s disease.

Table 1

Diagnostic data

Visit No.	Symptom Code	Disease Code
T1	S1, S2	D1
T2	S1, S2, S3	D1, D2
T3	S1, S3	D3
....	...	...

(2) Construction of Disease Diagnostic Model

Due to the differences in the experience and technical level of different doctors, some doctors might give a wrong diagnosis. A disease diagnostic model is constructed with the relationship between the patient symptoms and disease being used to determine the abnormal condition of the wrong treatment. Among them, the key method used is the association rule. The specific modeling steps are as follows:

Step 1: Find the historical medical history of the relevant disease as a training set and establish a medical data list;

Step 2: Generate a frequent item set using the Apriori algorithm. The specific process is to set the minimum support s and the minimum confidence c, scan the data set of the treatment data table, and generate candidate item set C1. The support item count of the candidate item set was compared with the minimum support count to generate a frequent item set L1. The algorithm continued to use the progressive link and pruning steps until the final frequent item set Lk was generated.

Step 3: The frequent item set generated an association rule. The specific process is to generate a non-void subset of Lk after generating the final frequent item set Lk. The non-void subset used the following formula:

$$confidence\ (S \Rightarrow D) = \frac{\sup port\ \_\ count(SUD)}{\sup port\ \_\ count(S)} \geq \min\_\ conf \tag{1}$$

After calculating whether the support is greater than the minimum confidence, an association rule can be generated. The disease type D can be derived from the symptom S, and finally the doctor diagnostic results are compared for judgment.

2.3.2 Excessive Treatment Discovery Method

Over-medication is the main manifestation of excessive treatment. By constructing a similar drug comparison model, this study has compared the similarity between new prescription and the best expert prescription to examine whether it is over-medicated with a new prescription. In order to facilitate the mining of the model, the specific prescription information was defined in a uniform, formal and abstract manner, as shown in Table 2.

Table 2

Formal and abstract definitions

Prescription information	Symbol
Collection of therapeutic functions of prescribed drugs	Fun_s
Statistics on the occurrence of therapeutic functions	w
Expert prescription	pre_j
Best expert prescription	pre_{best}
New prescription	pre_{new}
Collection of the same drugs for which best prescription and new prescription are available	pre_{pro}
Drug function similarity	Sim_{fun}
Classes of similar drugs	k_{sim}
Similarity between prescriptions	Sim_{last}
Threshold of prescription similarity	$threshold_{sim}$
Fun_z therapeutic functions of expert prescription drug A_i	A_i-fun_z
Fun_z therapeutic functions of new prescribed drug B_j	B_j-fun_z
The number of the same drugs for best and new prescriptions	num
The number of drug types for best prescription	h

(1) Best Expert Prescription

The best expert prescription should be determined before the similar drug comparison model was constructed. Assuming that in the prescription diagnosed by the doctor, the disease D had m symptoms, and the expert prescription was used to find the (N>0) disease expert prescription through the disease D. Secondly, according to the symptoms, the n(1≤n≤N) expert prescriptions most similar to the m symptoms were found. Among them, n expert prescriptions were the prescriptions with the highest matching degree similar between the symptoms in the prescription and m symptoms. If n=1, the prescription was the best expert prescription; otherwise, the “Treatment Weight Effect Approach” would be used repeatedly until the best expert prescription was found.

The basic idea of the “Treatment Weight Effect Approach” is: For the n similar expert prescriptions found, when the drug in the expert prescription has a certain therapeutic function, w=w+1 (the initial value is 0). Provided that wi denotes the count of therapeutic functions funi, if the drug of expert prescription prej has therapeutic function funi, its weight wji=wi; otherwise, wji=0. Then, the total weight Wj of the prescription prej was calculated, and the final total weight of all n prescriptions was compared; the prescription with the largest weight was the prebest. If there were multiple prescriptions with the same maximum total weight, then the least prescription cost of these prescriptions was selected as the best prescription. Table 3 shows a two-dimensional table for calculating the final total weight.

Table 3

Calculation of expert prescription treatment weight

Therapeutic functions **Prescription**	**fun$_1$**	**fun$_2$**	**...**	**fun$_k$**	**Total weight**
pre$_1$	w_{11}	w_{12}	...	w_{1k}	W_1
pre$_2$	w_{21}	w_{22}	...	w_{2k}	W_2
...	...	...	...	...	...
pre$_n$	w_{n1}	w_{n2}	...	w_{nk}	W_n

(2) Construction of Similar Drug Comparison Model

The best expert prescription was found and a drug comparison model was constructed to determine the abnormal treatment anomalies. There are five main steps in model construction. Among them, if prebest={A1,A2,...,Ah}, then new prescription is prenew={B1,B2,...,Bl}。

Step 1: Check whether pre$_{new}$ has any drug that is incompatible with the disease; if yes, it is judged that the prescription is wrong and invalid; if not, continue Step 2.

Step 2: Define Sim$_{first}$ as the initial similarity, indicating the similarity of pre$_{new}$ and pre$_{best,}$ or the exact same drug between prescriptions. Sim$_{first}$ is calculated according to the formula below:

$$Sim_{first} = \left| \frac{pre_{new} \cap pre_{best}}{pre_{best}} \right| \tag{2}$$

When Sim_{first} is 1, pre_{new} and pre_{best} are exactly same, which exclude the excessive treatment. When $pre_{best} \neq pre_{new}$, three conditions need to be considered:

(1) If $pre_{best} \cap pre_{new} = pre_{new}$ and $pre_{best} >> pre_{new}$, pre_{new} is considered to have the possibility of under-treatment.

(2) If $pre_{best} \cap pre_{new} = pre_{best}$ and $pre_{new} >> pre_{best}$, pre_{new} is considered to have a possibility beyond the actual diagnosis and treatment needs.

(3) In addition to the above two cases, if there is a drug different from pre_{best} in pre_{new}, the similarity of pre_{new} and pre_{best} cannot be judged by the drug name at this time. Continue Step 3, judge the similarity between drugs by using therapeutic functions, further judge the similarities between prescriptions, and finally identify the excessive treatment problem.

Step 3: Based on Hamming distance, the overall similarity of the therapeutic function is calculated, and then the similarity of the two drugs is judged. The specific calculation formula should be improved accordingly. K-Modes algorithm uses Hamming distance. Comparing the attribute values of sample data and cluster center, the number of different attribute values is the distance between them. Among them, if an attribute value is different, the distance is increased by 1, otherwise it is increased by 0. The model is designed to improve in the following four ways:

(1) If A_i-fun_z=0 and B_j-fun_z=0, both drug A_i and drug B_j are considered to have the same therapeutic function with a distance of zero.

(2) If A_i-fun_z=0 and B_j-fun_z=1, the drug A_i is considered to have this therapeutic function, but the drug B_j is not, and the distance is 1.

(3) If A_i-fun_z=1 and B_j-fun_z=0, the drug B_j is considered to have this therapeutic function, but the drug A_i has not, the distance is -1, indicating that the drug B_j may have a better therapeutic effect than the drug A_i.

(4) If A_i-fun_z=1 and B_j-fun_z=1, both drug A_i and drug B_j are considered to have no such theoretical function, and the distance is 0.5, indicating that drug B_j and drug A_i may have a reduced therapeutic effect on the symptom.

They are represented by the following two-dimensional table:

Table 4

Two-dimensional distance table for similarities of therapeutic functions similarity

Therapeutic functions / **Drug**	**fun_1**	**fun_2**	**fun_3**	**fun_4**	**...**
A_i-fun	0	0	1	1	...
B_j-fun	1	1	0	1	...
distance	0	1	-1	0.5	...

Then, the theoretical function distances are added and the formula is:

$$Sim_{fun} = \sum_{i=1}^{m} dis\tan ce_i \tag{3}$$

The smaller the Sim_{fun}, the larger similarity between pre_{new} drug B_j and pre_{best} drug A_i ; vice versa, the smaller the similarity is. If Sim_{fun} is less than a certain threshold, for example Sim_{fun} is 1, both drugs A_i and B_j are considered to be similar drugs.

Step 4: Calculate the similarity between pre_{new} - pre_{pro} drug B_j and $pre_{best\text{-}}$$pre_{pro}$ drug A_i, among which pre_{new} - pre_{pro} represents the difference set. In case B_j is similar to A_i, then $k_{sim}=k_{sim}+1$, until all the different functions of the collection drug are completed. Finally, the similarity with the final expert prescription Sim_{last} is calculated as:

$$Sim_{last}=(num+k_{sim})/h \tag{4}$$

Step 5: If $Sim_{last} \geq threshold_{sim}$, the similarity between pre_{new} and pre_{best} is deemed to be high. In other words, pre_{new} has better therapeutic effects over the disease and excessive treatment is excluded.

Through the above five steps, the construction of the similar drug comparison model is completed, and the excessive treatment can be distinguished from the perspective of prescription medication. Then the costs of prescriptions are compared, even if the prescription medication has no excessive treatment anomalies, to check whether the pre_{new} fee is higher than the pre_{best}. It can be calculated and determined by this formula:

$$\frac{F_{new} - F_{pro}}{F_{pro}} \times 100\% \leq threshold_{otal} \tag{5}$$

Where, F_{new} is the total cost of new prescription, and F_{pro} is the total cost of expert prescription. $Threshold_{total}$ represents the threshold of total costs, which are determined by the research of relevant institutions. If the total cost of pre_{new} does not exceed the $threshold_{total}$ of total cost, then there is no excessive treatment.

3. Results and discussion

This part includes experimental verification and discussion of the results.

3.1 Experimental Verification

In this part, erroneous and excessive treatment diagnostic issues are verified experimentally.

3.1.1 Experiment for Erroneous Treatment

According to the erroneous treatment discovery method, data was used for experimental verification. First of all, we preprocess the data, then use the model to complete the experiment, and finally summarize the results.

(1) Preprocessing of Experimental Data

In this part, 396 cases of historical diagnosis and treatment prescriptions stored in a hospital with expert prescription knowledge base were used as training set data to conduct association mining analysis and carry out experiments to verify the erroneous treatment problem. Due to the complexity of the medical data and the different formats, medical data needs to be pre-processed. A total of 16 major symptoms and 6 diseases were extracted from these 396 samples. The data processing is shown in Table 5, Table 6, and Table 7.

Table 5

Comparison table of symptom codes

Code	Symptom	Code	Symptom
S1	Cough	S9	Low fever
S2	Nasal congestion	S10	High fever
S3	Running nose	S11	Loss of appetite
S4	Stomach ache	S12	Weakness
S5	Headache	S13	Poor breathing
S6	Dizziness	S14	Pale skin
S7	Sore throat	S15	Fatigue
S8	Thinness	S16	Nausea and vomiting

Table 6

Comparison table of disease codes

Disease code	Disease name
D1	Viral influenza
D2	Influenza
D3	Pneumonia
D4	Iron deficiency anemia
D5	Hepatitis
D6	Gastritis

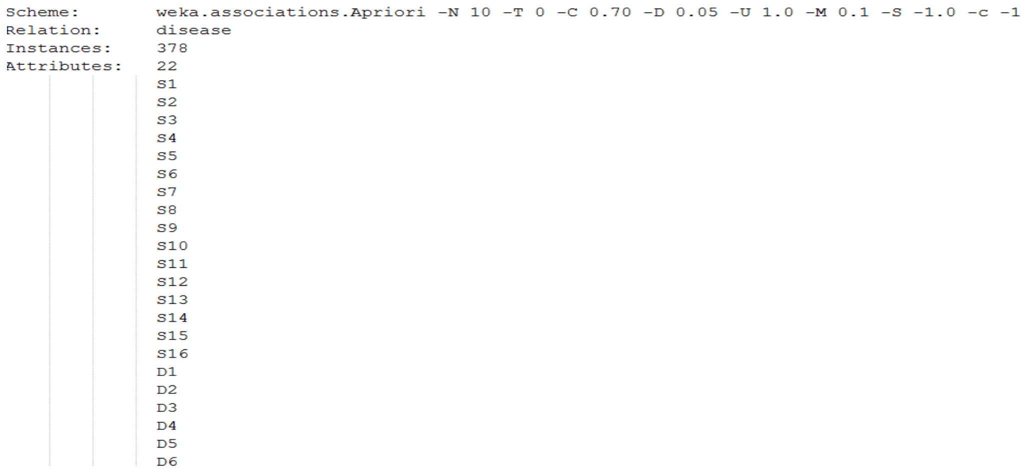
Table 7

Table of Visits (Transactions)

Visit Code	Symptom Code	Disease Code
T1	S1, S2, S3, S5, S6	D1
T2	S2, S3, S6	D1
T3	S1, S2, S5, S10	D2
...	...	...
T396	S1, S2, S5, S6, S9	D3

(2) Construction of disease diagnostic model association rule

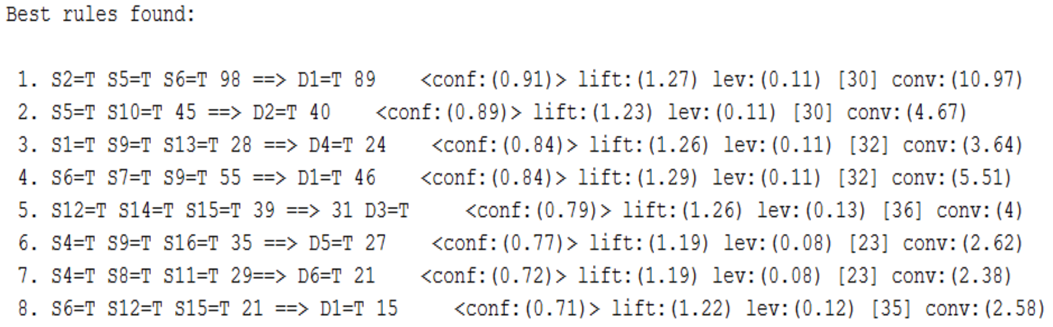
In the experiment, the association analysis function of weka software was used to perform the association rule mining for the disease diagnosis, and the data of Table 7 was written into the file of the disease.arff (the format of the weka file). According to the actual situation, the minimum support level set by the medical research institution is min_sup=0.05 and min_conf=0.70. The training attribute setting diagram of the weka association rule is shown in Fig. 2.



```
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.70 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     disease
Instances:    378
Attributes:   22
              S1
              S2
              S3
              S4
              S5
              S6
              S7
              S8
              S9
              S10
              S11
              S12
              S13
              S14
              S15
              S16
              D1
              D2
              D3
              D4
              D5
              D6
```

Fig. 2 Training attribute setting diagram for the weka association rule

The minimum confidence needs to reach 0.70 or more, so the training outputs 8 associative rules that meet the conditions, as shown in Fig. 3.



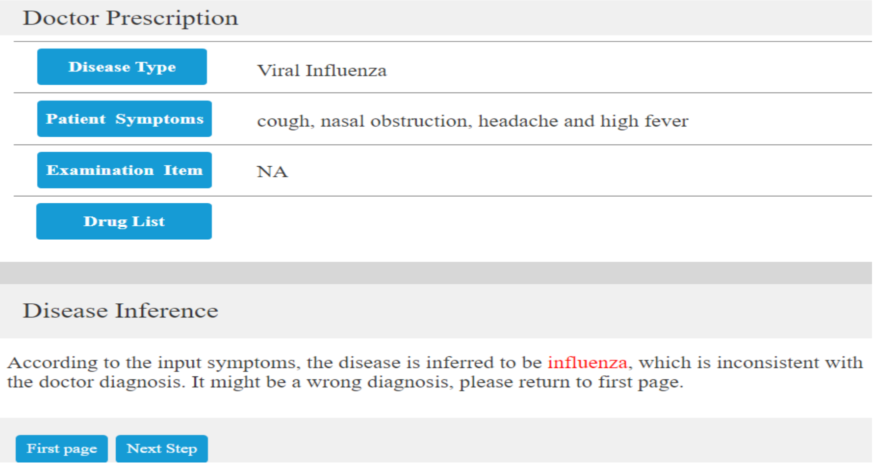
```
Best rules found:

 1. S2=T S5=T S6=T 98 ==> D1=T 89    <conf:(0.91)> lift:(1.27) lev:(0.11) [30] conv:(10.97)
 2. S5=T S10=T 45 ==> D2=T 40    <conf:(0.89)> lift:(1.23) lev:(0.11) [30] conv:(4.67)
 3. S1=T S9=T S13=T 28 ==> D4=T 24    <conf:(0.84)> lift:(1.26) lev:(0.11) [32] conv:(3.64)
 4. S6=T S7=T S9=T 55 ==> D1=T 46    <conf:(0.84)> lift:(1.29) lev:(0.11) [32] conv:(5.51)
 5. S12=T S14=T S15=T 39 ==> 31 D3=T     <conf:(0.79)> lift:(1.26) lev:(0.13) [36] conv:(4)
 6. S4=T S9=T S16=T 35 ==> D5=T 27    <conf:(0.77)> lift:(1.19) lev:(0.08) [23] conv:(2.62)
 7. S4=T S8=T S11=T 29==> D6=T 21    <conf:(0.72)> lift:(1.19) lev:(0.08) [23] conv:(2.38)
 8. S6=T S12=T S15=T 21 ==> D1=T 15     <conf:(0.71)> lift:(1.22) lev:(0.12) [35] conv:(2.58)
```

Fig. 3 Output of association rule training results

From the rules {S1,S5,S6}=>{D1}, the probability of a patient with viral influenza (common cold) is 91% estimated by the symptoms of nasal congestion, runny nose and dizziness (89 valid cases out of 98 cases). It can be inferred from {S5,S10}=>{D2} that the main symptoms are “high fever”, but the reasoned disease cannot be considered as a viral influenza but an influenza, which is consistent with daily medical experience. If the patient has a high fever (body temperature >38.2℃) and other symptoms of the common cold, the patient is usually suffering from the flu. In general, the results of erroneous medical judgments are applicable; for atypical diseases such as chronic diseases, it is necessary to re-diagnose according to the actual etiology.

(3) Experimental verification result

The experimental platform collected the doctor diagnosis prescription of a patient, and performed the worst treatment judgment according to the disease inference rule, as shown in Fig. 4.



Doctor Prescription

Disease Type	Viral Influenza
Patient Symptoms	cough, nasal obstruction, headache and high fever
Examination Item	NA
Drug List	

Disease Inference

According to the input symptoms, the disease is inferred to be influenza, which is inconsistent with the doctor diagnosis. It might be a wrong diagnosis, please return to first page.

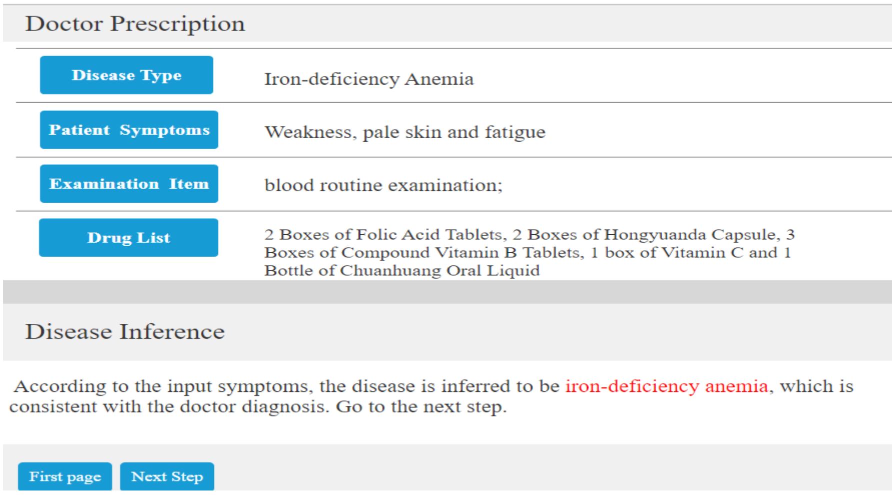
Fig. 4 Disease diagnosis

3.1.2 Experiment for Excessive Treatment Problems

According to the excessive treatment discovery method, it is applied in the experimental system. The experimental system completes the experiment according to the model, and displays the results. The following is the specific experimental process.

(1) Discovery of best expert prescription

The experiment is conducted with a new prescription from a hospital patient, as shown in Figure 5. The drug set is Medicine$_{new}$={“Hongyuanda Capsule”, “Folic Acid Tablets”, “Compound Vitamin B Tablets”, “Vitamin C Tablets”, “Zhenyuan Capsule”, “Chuanhuang Oral Liquid”}, and there is no Erroneous treatment.



Doctor Prescription

Disease Type: Iron-deficiency Anemia

Patient Symptoms: Weakness, pale skin and fatigue

Examination Item: blood routine examination;

Drug List: 2 Boxes of Folic Acid Tablets, 2 Boxes of Hongyuanda Capsule, 3 Boxes of Compound Vitamin B Tablets, 1 box of Vitamin C and 1 Bottle of Chuanhuang Oral Liquid

Disease Inference

According to the input symptoms, the disease is inferred to be iron-deficiency anemia, which is consistent with the doctor diagnosis. Go to the next step.

Fig. 5 Patient prescription

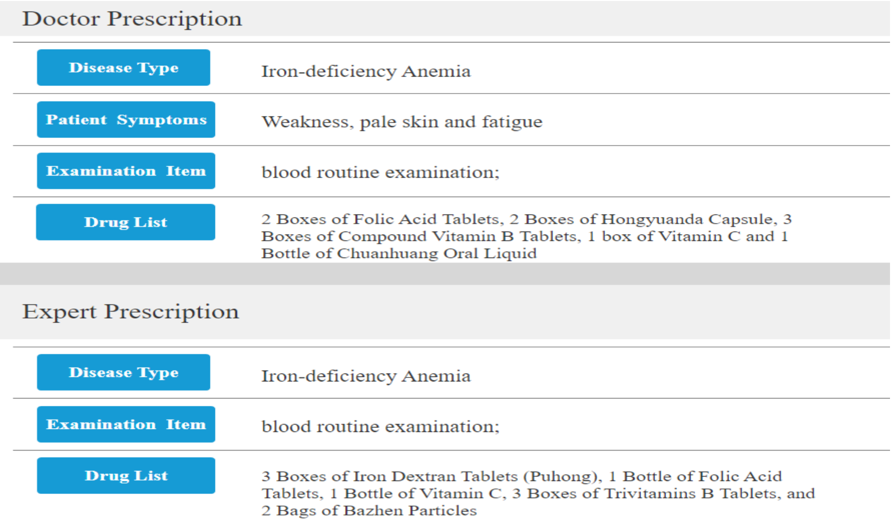
In this study, all the 358 expert prescriptions were identified by the disease “iron-deficiency anemia”, and the most relevant 53 expert prescriptions were found according to their symptoms. The calculation was conducted through the “Treatment Weight Effect Approach”, and the results are shown in Table 8.

Table 8

Calculation of expert prescription therapeutic weights

Function / **Prescription**	**Iron supplement**	**Vitamin B**	**Vitamin C**	**Dyspepsia**	**Dizziness and headache**	**Hemoglobin and red blood cells**	**Weight**
pre_1	52	0	41	35	30	23	181
pre_2	52	35	41	35	30	23	216
...	...	...	...	...	...	...	...
pre_{53}	52	35	41	35	0	0	163

The final total weights of the five expert prescriptions were the largest and the same, and the expert prescription with the lowest total cost was selected as the pre_{best} of iron-deficiency anemia. The result is shown in Fig. 6.



Doctor Prescription

Disease Type	Iron-deficiency Anemia
Patient Symptoms	Weakness, pale skin and fatigue
Examination Item	blood routine examination;
Drug List	2 Boxes of Folic Acid Tablets, 2 Boxes of Hongyuanda Capsule, 3 Boxes of Compound Vitamin B Tablets, 1 box of Vitamin C and 1 Bottle of Chuanhuang Oral Liquid

Expert Prescription

Disease Type	Iron-deficiency Anemia
Examination Item	blood routine examination;
Drug List	3 Boxes of Iron Dextran Tablets (Puhong), 1 Bottle of Folic Acid Tablets, 1 Bottle of Vitamin C, 3 Boxes of Trivitamins B Tablets, and 2 Bags of Bazhen Particles

Fig. 6 New prescription and expert prescription

(2) Verification of similar drug comparison model

Because there is no incompatible drug in the prescription, similarity can be compared between prescriptions. Also, the excessive treatment can be judged from the perspective of prescription medication. First of all, $Sim_{first}=|pre_{new} \cap pre_{best}|/|pre_{best}|=2/5=0.4$, namely, pre_{new} and pre_{best} have two similar drugs. Secondly, find similar drugs and find out the two groups of drugs of the same category - “Hongyuanda Capsule” and “Iron Dextran Tablets” and “Trivitamins B Tablets” and “Compound Vitamin B Tablets” - by adopting the drug classification method. Here, the distance threshold of each group of drugs is set to 1. The experimental process of similar drug search is as follows:

1) According to the drug knowledge base, it can be found that the “Hongyuanda Capsule” and “Iron Dextran Tablets” drugs have a relatively simple therapeutic function, and the effects are “iron supplementation” and treatment over iron-deficiency anemia. After calculation, distance = 0 < 1. Therefore, the two drugs are similar, and the number of similar drugs is k = 0+1 = 1.

2) The set of therapeutic functions for “Trivitamins B Tablets” and “Compound Vitamin B Tablets” is shown as Fig. 7:

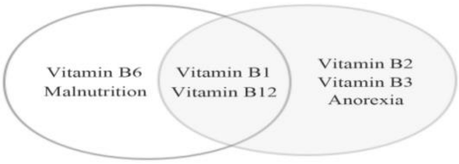


Fig. 7 Set of therapeutic functions

The main function of these two drugs is to supplement Vitamin B, and their other functions are calculated using Hamming distance to further discriminate the similarity. The similarity drug distance table is shown in Table 9:

Table 9

Distance of Drug Similarity

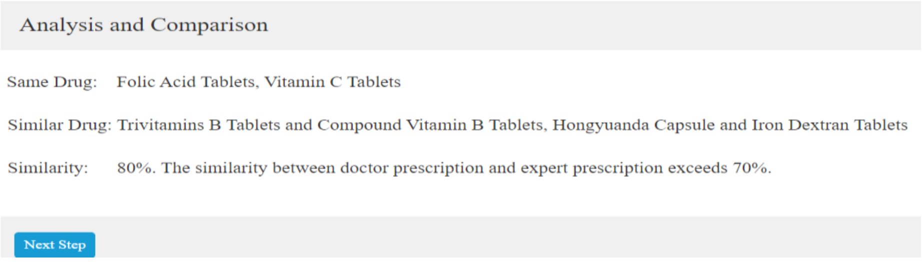
Therapeutic functions Drug	Vitamin B1	Vitamin B2	Vitamin B3	Vitamin B6	Vitamin B12
Trivitamins B tablets	0	1	1	0	0
Compound vitamin B Tablets	1	0	0	1	0
Distance	0	-1	-1	1	0

Sim_{fun} = 0+(-1)+(-1)+1+0 = -1 < 1, so these two drugs are similar, and k = k+1 = 2

3) For Bazhen Particles and Chuanhuang Oral Liquid, since the two drugs are not the same type of drug, it can be considered that the two drugs have no similarity. As a result, Sim_{last}=(num+k)/h=(2+2)/5=0.8, namely, the final similarity between new prescription and best prescription is 80%.

(3) Experimental Verification Conclusion

The drug's similarity threshold set in this study is 70%, so this new prescription is similar to the best expert prescription and is suitable for the treatment of iron-deficiency anemia with no excessive treatment. Among them, the threshold is selected according to the actual situation. The experimental results are shown in Fig. 8:



Analysis and Comparison

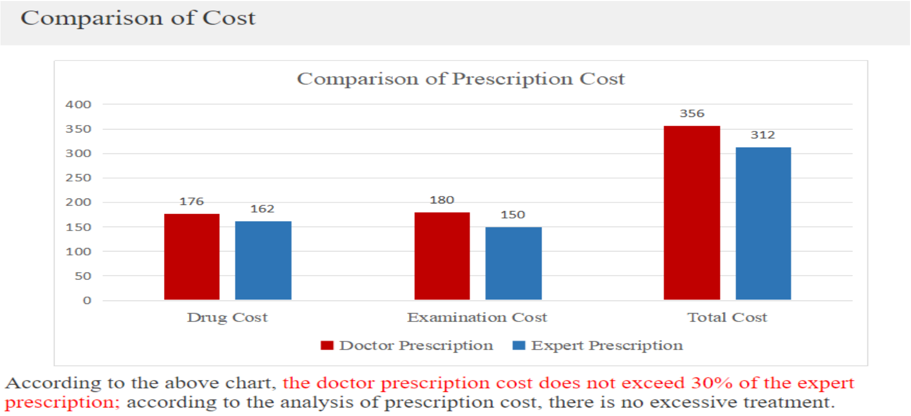
Same Drug: Folic Acid Tablets, Vitamin C Tablets

Similar Drug: Trivitamins B Tablets and Compound Vitamin B Tablets, Hongyuanda Capsule and Iron Dextran Tablets

Similarity: 80%. The similarity between doctor prescription and expert prescription exceeds 70%.

Fig. 8 Prescription similarity results

In order to determine whether there is an excessive treatment for the drug cost, the prescription cost is compared for the final step. From Fig. 9, the total cost of the doctor prescription is 356 yuan, whereas the total cost of the expert prescription is 312 yuan. According to the formula, ((356-312)/312)×100%=14%<30% can be obtained, so it can be considered that there is no excessive treatment in the new prescription.



According to the above chart, the doctor prescription cost does not exceed 30% of the expert prescription; according to the analysis of prescription cost, there is no excessive treatment.

Fig. 9 Comparison of prescription cost

3.2 Conclusions and Research Prospects

In this paper, we have reported our related research program being carried out on erroneous and excessive treatment anomalies. The association process and Hamming distance knowledge of unsupervised learning method were used to supervise the medical diagnosis and treatment process. Our study focused on whether there is an erroneous disease diagnosis during the doctor diagnosis and treatment, whether there was over-medication and excessive costs. Based on the massive medical data resources, we have constructed a disease diagnostic model, explored the relationship between disease and symptoms and identified the erroneous treatment problems. In the meanwhile, a “Treatment Weight Effect Approach” was employed in the prescription resource and the best expert prescription was found from multiple expert prescriptions. Subsequently, our study obtained the similarity between the drugs using the similar drug comparison model, thereby getting the similarity of the doctor prescription and the best expert prescription as well as judging the excessive treatment from the perspective of medication. In addition, the comparison of drug costs also helped to detect or rule out excessive treatment anomalies. Relevant cases of diagnostics and treatments were used in our experiment for verification analysis during an accurate verification process. The results obtained have shown certain practical significance. The experiment showed that the model implementation provides patients with a safe secure diagnosis and treatment environment, allowing effective recommendations for doctors diagnosis and treatment. Due to limited experimental conditions, the test data set is relatively small, and the effectiveness of the experiment also needs to be verified by collecting feedback from health experts such as doctors. The method has certain applicability in erroneous medical diagnosis, but secondary verification is needed for atypical cases. In the excessive medical diagnosis, the threshold of drug similarity needs to be set according to the actual situation.

In future research, we plan to further expand to much larger data sets to validate the accuracy of the abnormal detection of erroneous and excessive treatments. Another future expansion is to make the preliminary treatment of the medical information more standardized. Further research may combine some ideas of deep-learning algorithms, so as to solve and deal with erroneous and excessive treatment anomalies in the medical diagnosis and treatment process in a faster and more precise way.

Acknowledgements

Study is supported by Guangdong Science and Technology Program, China (No. 20140401).

REFERENCES

[1] Rowe I. A. (2018). Too Much Medicine: Overdiagnosis and Overtreatment of Non-alcoholic Fatty Liver Disease, Lancet Gastroenterology & Hepatology, 3(1), 66-72.

[2] Bianchine J. R. (1972). Drugs—Use and Misuse, Pastoral Psychology, 23(7), 56-60.

[3] Kalra J., Kopargaonkar A (2017). Quality Care and Patient Safety: Strategies to Disclose Medical Errors, Proceeding of the International Conference on Applied Human Factors and Ergonomics, Springer, Verlag, 159-167.

[4] X. L. Li, Q. X. Zhang, Y. W. Wang, et al (2017). Analysis and Explanation of Medication Errors Big Data in Chinese Medical Institutions, Chinese Journal of Pharmacoepidemiology, 1(26), 40-45.

[5] Stephanie M. (2019). My Mammogram Waiting Game, Mother Jones, 44(2), 70-71.

[6] Westfal M. L., Perez N. P., Hung Y. C., et al (2019). Pathologic Discordance to Clinical Management Decisions Suggests Overtreatment in Pediatric Benign Breast Disease Breast Cancer Research and Treatment, 176(1), 101-108.

[7] S. Q. Liu, S. Tang, Zhao J. F., et al (2019). Anetendedtopic Model Based Abnormal Medical Prescription Detection Method, Frontiers of Computer Science and Technology, 1-2.

[8] Z. J. Wei, T. Jin, J. M. Wang (2018). Outlier Detection Method in Healthcare Process Based on Clinical Data Mining, Computer Integrated Manufacturing Systems, 24(7), 1631-1632.

[9] Gelatti G. J., Carvalho A. C. P. D. L. F. D., Rodrigues P. P. (2017). Anomaly Detection through Temporal Abstractions on Intensive Care Data: Position Paper, Proceeding of the IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), Washington, D. C. , USA, IEEE, 354-355.

[10] Tago K., Jin Q. (2018). Detection of Anomaly Health Data by Specifying Latent Factors with SEM and Estimating Hidden States with HMM, Proceeding of the IEEE 9th International Conference on Information Technology in Medicine and Education. Washington, D. C., USA, IEEE, 137-141.

[11] C. K. Zhang, Y. Y. Chen, A. Yin, et al (2019). Anomaly Detection in ECG Based on Trend Symbolic Aggregate Approximation, Mathematical Biosciences and Engineering, 16(4), 2154–2167.

[12] Y. Liu, Z. Zhang, Q. L. Zhou (2014). Research on Fuzzy Rough Sets Based Rule Induction Methods for Healthcare Data, Computer Science, 41(12), 164-167.

[13] H. C. Song, Z. Q. Jiang, A. D. Men, et al (2017). A Hybrid Semi-Supervised Anomaly Detection Model for High-Dimensional Data, Computational Intelligence and Neuroscience, (1), 1-9.

[14] Rush B., Celi L. A., Stone D. J. (2018). Applying Machine Learning to Continuously Monitored Physiological Data, Journal of Clinical Monitoring and Computing, (1), 1-7.

[15] Goldstein M., Uchida S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data, Plos One, 11(4), 1-31.

[16] Amor L. B., Lahyai I., Jmaiel M., (2019). AUDIT: AnomaloUs Data Detection and Isolation Approach for Mobile Healthcare Systems, Expert Systems, (1), 1-21.

[17] Cai J. J. (2015). The Research and Implementation on Collaborative Treatment Platform based on E-Health, Shantou City, Shantou University, 2015.

[18] W. Liu, et al (2012). "E-Healthcare Interconnection Infrastructure Challenges and Solutions Overview”, Proceedings of HealthCom-2012, International Conference on E-Health Communications, Services and Applications, in Qing Hua University, Beijing, China.